基于 GIS 的河流水体污染非线性预测系统研究*

谭钦文[1,2]， 尹光志[1,2]， 李东伟[1,2]

(1. 重庆大学 资源及环境科学学院，重庆 400044；2. 重庆大学 西南资源开发及环境灾害控制工程教育部重点实验室，重庆 400044)

摘要：针对河流水体污染物的空间分布特点，提出综合运用 GIS、BP 神经网络和遗传算法，实现河流水污染的空间数据管理和污染预测的方法。该方法通过改进激励函数、为权值的修正加入动量项等方法改良 BP 算法；并引入遗传算法实现 BP 神经网络隐层节点数、最佳学习率和动量因子等参数的自动搜索，有效地解决了传统模型参数难以确定等问题。并进一步将该模型与 GIS 强大的空间功能结合，实现了水体污染的海量空间数据管理及评价预测结果的空间图形直观可视化表达，十分便于及时掌握河流水体污染动态、空间分布及演化趋势。并最终以 GIS 为二次开发平台，实现了基于遗传神经网络的河流水体污染非线性预测管理系统，并在长江重庆城区段河流污染预测应用中显示出良好的效果，预测精度达 78% 以上。

关键词：遗传算法；BP 人工神经网络；地理信息系统(GIS)；水体污染预测系统

中图分类号：X522 **文献标识码**：A **文章编号**：1006 - 7329(2006)05 - 0115 - 04

Nonlinear Forecast System for River Water Pollution Based on GIS

TAN Qin - wen[1,2], YIN Guang - zhi[1,2], LI Dong - wei[1,2]

(1. College of Resource and Environmental Science, Chongqing University, Chongqing 400044, P. R. China; 2. The Key Laboratory of the Exploitation of Southwest Resources & the Environmental Hazards Control Engineering, Ministry of Education, Chongqing University, Chongqing 400044, P. R. China)

Abstract: Based on the analysis of the water pollution spatial distribution characters of Yangtze River in Chongqing, a new method based on the integration of BP neural network and genetic arithmetic (GA) is proposed. For some shortcomings existed in the standard BP neural network, this method has ultimately overcome these shortcomings by combining the GA with BP artificial neural network through altering stimulating function, adding momentum factor to power value for BP algorithm and introducing genetic arithmetic to searching for the knots of the hidden layer, momentum factor and learning level. Using this method can easily overcome the difficulty of measuring the water prediction model's parameters. GIS is used as a tool for data management and spatial analysis, and the prediction result of the model for the water pollution spatial distribution characters of Yangtze River in Chongqing is visualized and explored with the precision of more than 78%.

Keywords: genetic arithmetic; BP artificial neural network, geographical information system (GIS); water pollution prediction system.

三峡大坝建成后，库区水体水质将如何变化一直是社会各界广泛关注的环境问题。面对社会各方压力，如何及时掌握库区水质污染及其空间分布状况，并对其现状及其变化趋势进行科学的预测评价是库区水污染管理迫切需要解决的问题。

水体污染评价对象作为一个复杂的非线性系统，不仅具有复杂的非线性结构，而且具有复杂的时空分布和演化特征，这给预测评价工作增加了不少困难。水质模型是研究和解决水体污染预测评价的一种常用方式，自从 1925 年美国的 Streeter 和 Phelps 导出 S - P 模型以来，各种水质模型得到了很大的发展[1]。但这些模型在应用中大多存在以下问题[2,3]：(1)模型前期

* 收稿日期：2006 - 03 - 20
基金项目：国家自然科学基金资助(编号 50374084)
作者简介：谭钦文(1976 -)，男，四川广安人，博士生，主要从事水环境技术和环境灾害控制工程的研究。

数据处理工作量大,模型建设周期长;(2)模型参数众多,结构复杂,模型检验和参数灵敏度分析工作量大;(3)受主观因素影响,人为误差较大;(4)模型对空间分布的模拟表达不够清晰和直观等。

近年来,遗传算法(Genetic algorithm,简称为GA)和BP人工神经网络(Error Back - Propagation Neural Network)方法在环境科学中得到了广泛的应用[4],加上GIS空间管理技术的飞速发展,为水体污染评价的最终解决提供了条件[2,3,5,6]。本文以长江重庆城区河流大渡口、朝天门和寸滩等三个不同断面的水环境监测统计资料为基础,建立了BP和GA相结合的水体污染非线性预测评价模型,有效的解决了传统建模参数难以确定、预测精度不高等问题,并进一步利用GIS来有效管理水体污染空间数据,并在空间分析的基础上直观表现污染的空间分布和变化规律及其影响范围。

1 预测系统整体技术构成

水环境信息具有复杂的空间分布特征,借助于地理信息系统(GIS)软件的空间管理和分析处理功能,把它的空间属性与属性特征结合起来进行一体化管理,可以有效地实现水环境管理的海量数据的输入、存贮管理以及模型分析处理结果的空间图形可视化表达与制图输出等与数据管理和表达有关的工作[2,3,6]。BP神经网络模型与遗传算法的结合可以有效地实现水质的非线性的、客观的评价分析[4,7~9]。研究中选用了美国ESRI公司的ARC/INFO8.3作为工作平台,并在此基础上使用VB进行二次开发,将GIS的空间管理与BP&GA的非线性预测功能进行有机结合,用来对河流水体污染进行预测评价管理,整个应用系统的结构体系如图1所示。

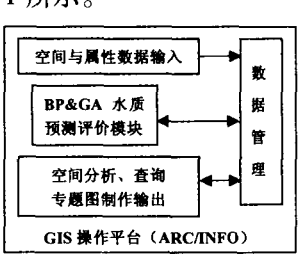

图1 预测系统整体流程图

2 BP&GA预测方法的确定

BP人工网络是目前在工程中应用比较广泛的一种人工网络[4,7],其核心是网络的误差反传。误差反传的主要思想是把学习过程分为两个阶段:第一阶段(正向传播过程),给出输入信息通过输入层,经隐含层逐层处理并计算每个单元的实际输出值。信息正向传播过程可由第k层第j个神经元的输入输出关系简单地表征为:

$$y_j^k = f_j^k\left(\sum_{i=1}^{n_k-1} W_{ij}^{(k-1)} \ominus y_i^{k-1} - \theta_j^k\right)$$

$$j = 1,2,\cdots,n;k = 1,2,\cdots,M$$

式中:$W_{ij}^{(k-1)}$为第$(k-1)$层第i个神经元到第k层第j个神经元的连接因子;θ_j^k为该神经元的阈值;$f(x)$为网络节点作用函数,本研究采用sigmoid函数;n_k为第k层神经元的数目,M为神经网络模型的总层数。

第二阶段(反向过程),若输出层未得到期望的输出值,则逐层递归地计算实际输出与期望输出之差(即误差),以便根据此误差调节权值。误差反传的算法有两种,一种是批处理算法,另一种是单个样本的训练方式。由于批处理方式更适合多样本的网络学习,所以,本文也采用批处理的误差反算法。修正系数采用梯度法:

$$\begin{cases} W(n+1) = W(n) - \eta \cdot \dfrac{\partial E}{\partial W(n)} \\ \theta(n+1) = \theta(n) - \eta \cdot \dfrac{\partial E}{\partial \theta(n)} \end{cases}$$

式中:η为网络学习因子。在网络训练过程中,为避免出现数值振荡,上式中常加上一动量项,引入动力因子α。另外由sigmoid函数的性质可知,神经元节点的输入绝对值太大时易出现神经元节点的饱和,因此对输入向量须作标准化处理。

由于BP神经网络自身存在的不可避免的缺点,例如:网络的结构、网络的学习率η、修正权值的附加动量项α,这三个至关重要的参数的确定并没有很好的办法,而遗传算法能很好地解决这个问题。所以本文关于河流水体污染预测采用BP人工神经网络与遗传算法相结合的方法。即:依靠遗传算法强大的搜索能力,搜索到适合本问题的最佳BP网络结构、学习率η、修正权值的附加动量项α,然后依靠BP网络的强大的学习能力,通过网络的训练,取得理想的环境污染的预测网络。

遗传算法的数学理论基础是Holland提出的图式(Schmata)定理[8,9]。图式是描述种群中任意染色体之间相似性的一组符号串。它由符号0,1,*定义,即由二进制数字0,1及通配符*任意组合而成。图式中0、1序列组成其固定部分,*表示其变化部分,整个图式表示有意义的匹配模式。由0,1,*定义,长度为L的符号串所能组成的最大图式数或相似性为$(2+1)^L$。若含有N个染色体的种群可能包含的图式数在$2^L \sim N \cdot 2^L$之间。遗传算法正是利用种群中包含的众多的图式及其染色体符号串之间的相似性信息进行启发式搜索和问题求解。已证明,在产生新一代的过程中,尽管遗传算法只完成了正比于种群长度N的计算

量,而处理的图式数却正比于种群长度 N 的三次方。

通过对遗传算法和 BP 人工神经网络的分析,可以肯定运用 GA 能有效防止搜索过程收敛于局部极小点、易于求得全局最优解和能并行搜索的特点,来很好解决用 BP 网络预测河流水体污染时三个重要参数(隐含层节点数目、学习率 η 和动量因子 α)难以解决的问题。

3　遗传算法和 BP 神经网络的结合

3.1　编码方式的确定

编码的恰当与否对问题求解的质量和速度有直接的影响,编码技术的研究成为当前遗传算法研究的重要分支。由于简单遗传算法采用二进制编码,虽然根据二进制的模式定理,知道采用二进制数编码比非二进数提供更多的图式,但在实际的运用中二进制编码有以下的缺点:处理的实际问题往往是十进制数,而用二进制数字编码时,需要把实际问题对应的十进制数转化为二进制数,使其数字长度扩大约 3.3 倍,这虽然可以扩大搜索域,但在输出结果时需要解码,编码再解码的工作量非常大,运算的效率较低,也可能使遗传算法的性能变坏。

许多学者通过研究十进制编码的模式定理指出:十进制整数编码的遗传算法的群体中模式的数目仅与群体大小和染色体长度有关,其中具有短的定义距、低阶并且适应度值在群体平均适应度值以上的模式在遗传算法迭代过程中将按指数增长率被采样。故本研究根据研究对象的具体特点,采用十进制编码方式。

3.2　初始种群生成与 GA 和 BP 的集成方式

初始染色体的多少对遗传算法的搜索和人工神经网络的性能都有一定的影响。染色体数目越多,训练结果的精度就越高,但所花的时间就越长[8,9]。为了加快优化速度,往往要对染色体参数加以适当的限制。针对本文问题,隐含层节点的数目在 1 ~ 200 之间,学习率 η 和动量因子 α 的值在 0.1 ~ 0.99 之间比较合适。初始种群的染色体数目可定为 4 条,染色体的结构如图 2 所示。

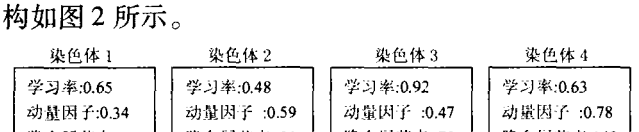

图 2　初始种群的染色体结构示意图

为便于遗传运算,染色体中的小数参数转化为整数,即学习率 η 和动量因子 α 为两位十进制整数,隐含层节点数为三位十进制整数,染色体长度为 7。整个 GA 和 BP 相结合预测流程如图 3 所示。

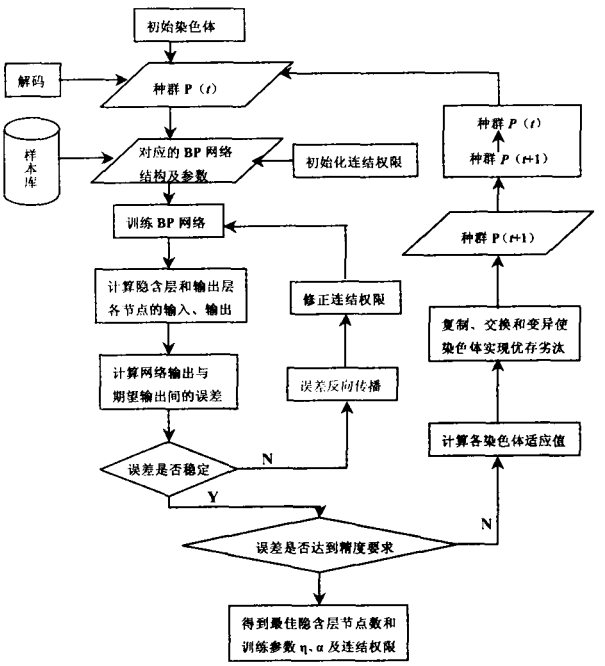

图 3　GA 和 BP 相结合预测流程图

4　预测系统的实现与应用

GA 和 BP 相结合实现水质非线性预测的设计目标是通过遗传算法自动搜索和样本库的训练来确定合理的 BP 神经网络结构和网络参数,从而可以运用合理的 BP 网络来对河流污染进行预测。根据 Kolmogrov 神经网络映射存在原理,一个三层前向神经网络能够实现任意连续函数的映射,所以本研究采用三层 BP 网络。根据研究对象的监测数据,取前一监测断面的常规监测项目:河水断面流速、溶解氧、SS、氨氮、COD 和断面间距等六项作为输入,来预测下一断面同一时间的溶解氧、SS、氨氮、COD(以 CODcr 计)等的状态值,即 BP 网络的输入结点取 6,输出结点为 4。在取定 BP 网络的层数、输入、输出结点数后,以研究区不同断面的监测数据为样本库,根据前面建立的 GA&BP 方法来搜索出最佳的网络结构参数为:隐含层节点数:27;最佳学习率:0.89;最佳动量因子:0.65。

将上述预测模型经 VB 编程与 GIS 集成实现后,由 GIS 数据管理模块向训练后的预测软件系统输入任意样本,本预测模型就得到对应的预测值、预测误差以及每次预测的误差直方图,并同时存贮于 GIS 数据库中以备进一步的分析处理所需。经运算得到相应输出节点的预测误差曲线图如图 4、5、6、7、8。

将地理配准的各监测点的预测结果经 GIS 叠加显

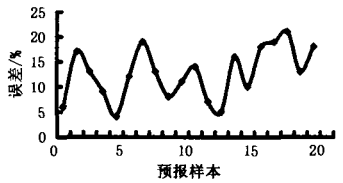

图4 输出节点1预测误差曲线

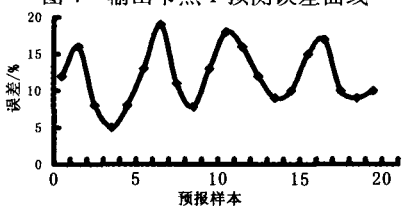

图5 输出节点2预测误差曲线

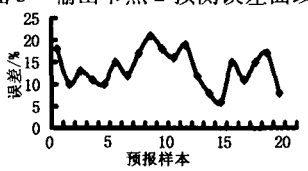

图6 输出节点3预测误差曲线

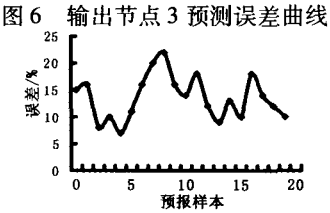

图7 输出节点4预测误差曲线

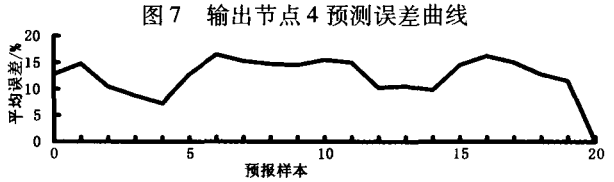

图8 预测误差曲线

示于河流水系及地形图形之上，制成相应的专题图件，将水质模拟的结果可视化地表达于计算机平台，经插值和缓冲区分析，研究区河流水质变化的时空特征和空间变化规律都可非常清晰直观地表现出来。

从以上误差曲线可以看出，预测误差都不超过22%，充分说明GA和BP神经网络的结合，所建立的河流水污染预测模型具有较高的预测精度，预测的结果具有相当高的可信度，且结果形象直观，十分有利于进一步挖掘现象背后隐藏的客观规律。

5 结论

由于标准BP神经网络算法自身存在易形成局部最小而得不到整体最优、训练易陷入瘫痪、收敛速度很慢、网络的泛化与推广能力比较差和隐含层节点难以确定等问题，本研究一方面采用改进激励函数和加入动量项等办法改良BP算法，另一方面通过引入遗传算法来搜索BP网络最佳结构参数，这样既克服了传统BP神经网络的缺点，又有效解决了河流水质预测评价模型参数确定的难题。同时将该模型与GIS技术相结合，充分应用GIS强大的空间数据处理能力，实现了水质管理中的海量空间数据管理及评价预测结果的空间图形直观可视化表达，并经初步应用证明该方法是可行的，它不仅具有预测精度高，数据管理方便的特点，而且以图形形式清晰直观呈现预测结果，十分有利于管理者及时掌握河流水体污染动态、空间分布及演化趋势，为进一步的决策分析提供便利。

参考文献：

[1] 傅国伟. 河流水质数学模型及其模拟计算[M]. 北京：中国环境科学出版社，1987.

[2] 马蔚纯，张超. 基于GIS的水质数值模拟－以上海市苏州河为例[J]. 地理学报，1998，(S0)：67－75.

[3] Parks Bradley O. The Need for Integration. In: Goodchild Michael F, Steyaert Louis T, Parks Bradley O. Environmental Modeling with GIS[J]. New York: Oxford Univ. Press. 193－93；30－34.

[4] 袁曾任. 人工神经网络及其应用[M]. 北京：清华大学出版社. 1999.

[5] 张行南，耿庆斋，逄勇. 水质模型与地理信息系统的集成研究[J]. 水利学报，2004，1：90－94.

[6] 何强，龙腾锐，夏志祥. 水污染控制系统规划方法研究[J]. 重庆建筑大学学报，1999，21(6)：31－34.

[7] 索胜军. 几种改进BP算法的性能比较[J]. 哈尔滨建筑大学报，2000，33(2)：80－83.

[8] 吴晓涛. 用遗传算法进行路径规划[J]. 清华大学学报，1995，33(5)：14－19.

[9] 顾峻. 遗传算法对模糊控制的优化及其应用[J]. 东南大学学报，1998，28(S0)：109－119.